第九章 异常检测算法

目录

9.1 异常检测算法概要

9.1.1 异常与异常检测

定义：

- 异常是与常规模式或预期行为不一致的数据点或活动。
- 异常检测就是对上述的异常情况进行分析或预警。

异常类型：

- **偏离点（Deviant）**：与预期结果显著差异的点（可能包含有价值信息）。
- **变化点（Change point）**：时间序列中的突变或转折点。
- **奇异模式（Surprise pattern）**：非典型的罕见模式。
- **新颖事件（Novel event）**：未曾观察到的新事件。
- **不一致（Disaccord）**：数据或信息中的矛盾。

异常检测分类：

- 监督学习：需标注数据。

- 半监督学习：部分标注数据。
- 无监督学习：无需标注数据（主流方法）。

主要难点：

1. 数据稀疏性（异常样本少）。
2. 噪声与异常值的区分。
3. 异常现象的解释困难。

总结：异常的定义因应用场景而异。例如，金融欺诈检测中的“异常”可能是高频交易，而工业检测中可能是设备振动异常。理解业务背景是定义异常的关键。

9.1.2 异常检测算法分类

异常检测算法的分类

类别	原理	异常判定标准	代表性算法
统计理论	假设数据服从某种分布，异常数据发生在低概率区域	偏离正常分布范围	3σ准则、箱线图、直方图、累积和法等
空间分布	异常值通常分布在大多数数据点距离较远或在密度较低的区域	数据点的相对距离大于阈值，或密度小于阈值时，为异常	孤立森林、局部异常因子等
降维	通过降维技术将数据从高维空间映射到低维空间。高维空间难以辨别，低维空间因数据分布简化凸显出来	根据数据在低维空间的分布情况（如密度）或者数据重构前后的误差来计算异常得分，超过阈值为异常	主成分分析、自编码器等
预测	利用历史数据构建模型，异常值与正常模式存在差异，预测值在异常值处会显著增大	实际观测值与异常值之间的差异超过预设阈值时，为异常	向量自回归模型、自回归差分移动平均模型、LSTM模型等

总结：统计方法适合简单场景，但需数据分布假设；空间分布方法对复杂分布更鲁棒；降维方法可解决高维问题，但需权衡信息损失；预测方法适合时序数据，但对模型复杂度敏感。

9.1.3 异常检测的常用数据集

9.1.3.1时间序列的异常检测

原理：识别X(abnormal)，且和X(normal)差异尽可能大，或者，X(abnormal)出现概率尽可能小

单变量时间序列

多变量时间序列

9.1.3.2图像异常检测

作用：识别那些与正常图像分布不一致的图像

步骤：

1. 输入图像I
2. 提取和比较图像特征F(I)
3. 计算其图像特征F(I)
4. 基于这些特征进行挖掘
5. 判断图像是否异常

主要应用领域：工业产品缺陷检测、医学影像分析、高光谱图像处理等

9.1.3.3视频异常检测

作用：识别不符合正常行为模式的视频片段

表现：突然的运动变化、外观变化或与其他的显著不一致

步骤：对视频序列进行逐帧分析——提取每帧特征——构建视频特征序列——比较特征序列与正常行为模式的差异——识别处异常行为所在的视频片段

9.2 基于统计理论的异常检测

9.2.1 3σ准则

假设D服从正态分布，数值分布在（μ-σ，μ+σ）（μ-2σ，μ+2σ）（μ-3σ，μ+3σ）的概率为68.3%、95.5%、99.7%，位于（μ-3σ，μ+3σ）区间外的为异常值

拓展：c·σ形式

算法步骤：

```
1．计算μ和σ
1．设定阈值（μ-3σ，μ+3σ）
1．判断异常在（μ-3σ，μ+3σ）区间外
```

优点：数据符合正态分布时表现优异，计算简单直观，能迅速识别出大部分异常值

局限性：依赖于正态分布特性，对非正态数据会产生误判，不够灵活无法适应异常情况

9.2.2 箱线图

四分位距 IQR=Q3-Q1

下截断点 CL=max{Xmin,Q1-1.5IQR}

上截断点 CU=max{Xmax,Q3+1.5IQR}

正常值X∈[CL,CU]，其他为异常值

拓展：c·IQR

算法步骤：

1. Xi由大到小排序
2. 计算Me、Q1、Q3
3. 计算IQR
4. 绘制箱线图
5. 判断异常

优点：处理不符合正态分布的数据时具有较高客观性；较强的稳健性

局限性：无法准确反映具有复杂或多模态分布数据的分布特征，会忽略某些异常行为

9.2.3 基于直方图的异常得分

HBOS：用于多变量时间序列的异常检测，基本假设时数据的各特征是独立的

原理：对数据集D中的每个特征Xi构建单变量直方图；对于分类数据，统计每个类别的频数并计算相对频率，作为直方图的高度；对于数值数据，使用静态宽度直方图和动态宽度直方图对连续数据离散化，计算概率密度

静态宽度直方图：划分柱形范围[Xmin+(k-1)·w,Xmax)，柱形高度hij=nij/k，各柱形宽度相同

动态宽度直方图：N/k个连续值为一组，共k组，将每组数绘制为一个柱形，各柱形面积相同

计算：数据点所处的柱形高度越低，其概率密度越小，异常概率越高，样本Xj的异常得分Score(Xj)=-log(P(Xj))

算法步骤：

1. 绘制直方图
2. 计算概率密度P(Xj)
3. 计算异常得分
4. 设定阈值ε，若Score(Xj)>ε

优点：处理速度快，具有线性时间复杂度，能够适应大规模数据集的处理需求

局限性：在高维度数据集中的检测结果较差；不同的区间划分对应着不同的密度函数，会影响异常检测的结果

9.2.4 累积和法（CUSUM）

- **原理**：监控子组均值的累积偏差，超过控制限则判定异常。
- **应用**：适合实时监控微小趋势变化（如质量控制）。

思考：统计方法简单高效，但需人工设定阈值，且对复杂模式（如非线性关系）适应性差。

9.2.5 实践案例：基于箱线图的 wiki 网络流量异常检测

利用箱线图对 wiki 网络流量进行异常检测，数据集来自 Kaggle 赛题。

经过数据预处理、模型构建、参数调整和结果可视化，发现中文 wiki 页面流量中位数低且有异常值，不同语言页面流量波动情况各异。

异常值不一定表示问题，可能由特定事件导致。

9.3 基于空间分布的异常检测

9.3.1 孤立森林（Isolation Forest）

原理：通过随机划分特征空间，异常值路径长度较短。
步骤：

1. 构建多棵孤立树。
2. 计算数据点的平均路径长度。
3. 根据路径长度计算异常得分。

评价：

- 优点：适合高维数据，计算效率高。
- 缺点：对噪声敏感。

9.3.2 局部异常因子（LOF）

原理：通过局部密度计算异常得分，密度越低得分越高。
步骤：

1. 计算k近邻距离和可达距离。
2. 计算局部可达密度和LOF得分。
3. 判断异常：LOF>1则为异常。

评价：

- 优点：适合不均匀分布数据。

- 缺点：计算复杂度高。

9.3.3 案例：基于LOF的信用卡欺诈检测

数据集：Kaggle信用卡交易数据（284,807笔，492笔欺诈）。
步骤：

1. 特征工程：时间转换为小时，删除冗余列。
2. 模型训练：`LocalOutlierFactor(n_neighbors=25)`。
3. 评估：Top N准确率分析。
   结果：

- 前22个样本准确率48%，前100个降至26%。
- 需结合其他方法提高准确性。

9.4 基于降维的异常检测

9.4.1 主成分分析（PCA）

原理：通过降维重构误差检测异常。
步骤：

1. 标准化数据并计算主成分。
2. 重构数据并计算误差。
3. 设定阈值判断异常。

评价：

- 优点：消除冗余特征，可视化方便。
- 缺点：对非线性关系处理能力弱。

9.4.2 自编码器（Autoencoder）

原理：通过编码-解码重构误差检测异常。
步骤：

1. 构建编码器和解码器。
2. 最小化重构损失（如MSE）。
3. 判断异常：误差超过阈值则为异常。

评价：

- 优点：捕捉非线性关系。
- 缺点：需大量训练数据。

9.4.3 案例：基于PCA的飞机引擎异常检测

数据集：引擎热量、振幅、转速等5个指标。
步骤：

1. 数据标准化。
2. PCA降维至3维。
3. 可视化异常得分。
   结果：

- 异常点在三维空间中明显偏离聚类区域。
- 紫色表示高异常概率，黄色为正常。

9.5 基于预测的异常检测

9.5.1 向量自回归模型（VAR）

原理：利用多变量时间序列的滞后项建模。
步骤：

1. 平稳性检验（ADF）。
2. 确定滞后阶数（AIC/BIC）。

3. 预测误差判断异常。

评价：

- 优点：捕捉动态关系。
- 缺点：计算复杂度高。

9.5.2 ARIMA模型

原理：结合自回归（AR）、差分（I）、移动平均（MA）。
步骤：

1. 差分使序列平稳。
2. 选择参数(p,d,q)。
3. 预测误差判断异常。

评价：

- 优点：处理趋势和季节性。
- 缺点：参数选择复杂。

9.5.3 LSTM

原理：通过门控机制捕捉长期依赖关系。
步骤：

1. 构建LSTM层和Dropout层。
2. 训练模型最小化MAE损失。
3. 预测误差判断异常。

评价：

- 优点：处理复杂时间模式。
- 缺点：训练时间长。

9.5.4 案例：基于LSTM的股票价格异常检测

数据集：标普500指数日收盘价（1986-2018）。
步骤：

1. 数据划分：训练集95%，测试集5%。
2. 归一化处理。
3. 构建LSTM模型（`units=64`）。
4. 设定阈值（前5% MAE）。
   结果：

- 异常点对应价格剧烈波动时段。
- 红色标记显示异常值。

9.6 延伸阅读——WSARE

原理：基于规则比较近期数据与基线分布的差异。
版本：

- **WSARE 2.0**：基线由过去两个月数据计算。
- **WSARE 3.0**：贝叶斯网络建模，考虑季节效应。

步骤：

1. 生成候选规则。
2. 计算规则得分（卡方检验）。
3. 随机化检验确定P值。

应用： 短期监测（如大型活动）使用WSARE 2.0；长期监测（如传染病）使用WSARE 3.0。

思考：WSARE结合规则与统计检验，适合领域知识驱动的场景，但实现复杂度较高。

总结：

方法选择：需结合数据特性（如维度、分布）、业务需求（实时性、解释性）和计算资源。

趋势：混合方法（如统计+机器学习）、自动化阈值调整、在线学习是未来方向。